Deep Learning for LiDAR Waveforms with Multiple Returns

Andreas Aßmann
EPS, Heriot-Watt University
STMicroelectronics R&D Ltd.
Edinburgh, UK
aa224@hw.ac.uk

Brian Stewart
STMicroelectronics R&D Ltd.
Edinburgh, UK

Andrew M. Wallace
EPS, Heriot-Watt University
Edinburgh, UK

***Abstract*—We present LiDARNet, a novel data driven approach to LiDAR waveform processing utilising convolutional neural networks to extract depth information. To effectively leverage deep learning, an efficient LiDAR toolchain was developed, which can generate realistic waveform datasets based on either specific experimental parameters or synthetic scenes at scale. This enables us to generate a large volume of waveforms in varying conditions with meaningful underlying data. To validate our simulation approach, we model a super resolution benchmark and cross-validate the network with real unseen data. We demonstrate the ability to resolve peaks in close proximity, as well as to extract multiple returns from waveforms with low signal-to-noise ratio simultaneously with over 99% accuracy. This approach is fast, flexible and highly parallelizable for arrayed imagers. We provide explainability in the deep learning process by matching intermediate outputs to a robust underlying signal model.**

***Index Terms*—LiDAR, Deep Learning, Convolutional Neural Networks, Super-Resolution, Time-of-Flight Imaging**

I. Introduction

Sensing depth from dynamic scenes is of crucial importance to enable reliable self driving cars, augmented and virtual reality and advanced driver assistance. High precision localisation of other objects surrounding the actor can be a safety critical task and requires fast and robust depth measurements at a fine spatial resolution. To achieve this task light detection and ranging (LiDAR) is often used. Photons are emitted via a laser source which are reflected back from objects in its path. Knowing the speed of light, one can determine the time between emission and detection and therefore determine the distance. In long range applications with a high dynamic range (e.g. 0-300 m at cm resolution), there is a high potential to encounter multiple returns from objects due to beam divergence and multiple reflections. This results in multiple returns in the LiDAR waveform; peaks from targets in very close proximity may merge for every single pixel or scan location of the Time-of-Flight (ToF) sensor. Further, ambient sunlight and adverse weather conditions will introduce non-surface returns. This makes it challenging to resolve and detect returns from the sensor as illustrated in Figure 1.

We are leveraging deep learning (DL) approaches to resolve closely separated surfaces and thus perform super-resolution on the LiDAR waveform with convolutional neural networks (CNNs) as well as extract multiple peaks in long range waveforms with high dynamic range. We design our neural networks based on exisiting LiDAR signal models and a sparse encoder stage with the hypothesis that we can trace model aspects throughout our network architectures. Our assumptions rely on an underlying sparse signal model [1] and practical signal aspects introduced by single photon counting sensing modalities [2]–[4] in array form. Our approachs provides a way to process a large volume of waveforms for large scale arrays simultaneously.

This work was supported by STMicroelectronics R&D Ltd. and the EPSRC Grant EP/L01596X/1.

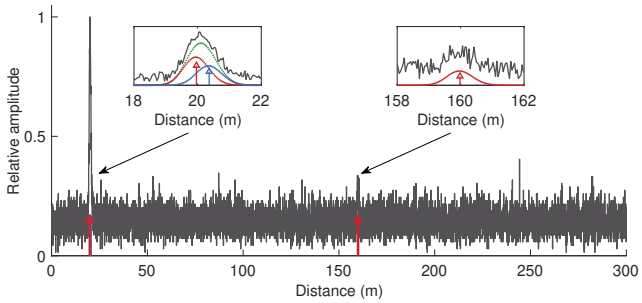


Fig. 1. LiDAR multiple return waveform problem over large dynamic range. Targets need to be individually resolved in proximity and other returns need to be detected with low SNR at long range.

Contributions of our work are summarized as follows:

- Signal conversion from noisy dense data to sparse peak locations via data driven machine learning approaches.
- Training by simulation: using a validated simulator with traceable activations, to eliminate need for large datasets.
- Performing super-resolution detection on a multi-return waveform using deep learning approaches.
- A deep learning architecture to perform fast peak (i.e. surface) localizations from complex LiDAR waveforms with multiple returns for large photon detector arrays.

To the best of our knowledge this is the first time DL has been applied to LiDAR waveforms to extract multiple returns. Before describing our signal model and simulation approach in section III, the network architecture with validation in section IV and results in section V, a brief overview of related work is provided.

II. Related work

The concept of sparsity is well understood and has been successfully adapted by the machine learning (ML) community as significant codes derived from high dimensional data. This principle is effectively used in so called auto-encoders [5], which have found useful applications in de-noising images and one dimensional signals, for example in [6], [7]. The same principle is exploited in many traditional signal processing methodologies including wavelet denoising [8], compressive sensing [9], compression or finite-rate of innovation deconvolution [1]. While there has been extensive work to process LiDAR waveforms using statistical and sparse methods [1], [10], [11], they are often designed for very specific conditions and will either require specific parameters for varying environmental conditions or only work in one single use case. Further, useful application of data driven methods are still to be explored for LiDAR waveforms. This is probably due to the lack of extensive full-waveform datasets, as most sensors disregard the full waveform and only store final depth measurements. Recent work in solid-state sensors for LiDAR applications and related modelling of the system behaviour provide valuable parameters to simulate LiDAR waveforms with real system parameters [3], [4], which are very good approximations of real waveforms. While ML has been applied to LiDAR waveforms before [12], it was only used to smooth and de-noise the waveform in the presence of minimal noise.

III. Signal Model and Data Generation

A. Signal model

We convert a complex LiDAR waveform to a sparse collection of peak locations with confidence values derived from ML. This is a large scale classification problem. We note that in practice there will be a finite amount of real peak locations due to the quantization in the sampling procedure. Previous work [1] has established that multiple returns of photons, h, can be modelled as a sparse signal, with each peak represented by a scaled Dirac distribution,

$$h(t) = \sum_{k=0}^{K-1} \Gamma_k \delta(t - t_k), \tag{1}$$

where δ is the Dirac distribution, $\{\Gamma\}_{k=0}^{K-1}$ denotes the strength of the k^{th} return and $\{t_k\}_{k=0}^{K-1}$ the respective time delay. We adapt this model for a classification problem with the only property of interest being the respective time delay t_k of each true surface return in form of an ideal binary signal. Hence, an ideal histogram $\mathbf{h} \in \{0,1\}^p$ can be described as,

$$\mathbf{h} = [0 \dots 1 \dots 1 \dots 0], \tag{2}$$

where p is the number of bins and each non-zero entry (equal to 1) is the location of a return at its respective discrete time delay closest to the quantized distance vector. Using ToF, $d = 2t_k/c$), this encodes the time delay with a bin resolution of $\pm d_{max}/2p$,

$$\mathbf{d} = [d_0 \dots d_p]. \tag{3}$$

This results in a sparse vector containing the peak locations such that

$$\mathbf{m} = [0 \dots 1 \dots 1 \dots 0] \circ [d_0 \dots d_u \dots d_p] \tag{4}$$

$$= [0 \dots d_k \dots d_{K-1} \dots 0], \tag{5}$$

where $\{d_k\}_{k=0}^{K-1}$ is the distance of each k^{th} return. This is an extension to the single surface return formulation in [13].

We consider the common single photon avalanche detector (SPAD) as a key component for time-correlated single photon counting (TCSPC) LiDAR. The sensing modality will have an influence on the final return signal shape for each underlying ideal impulse at location, k. One such trait is the slowly decaying tail, introduced by passive SPADs due to their dead time [2]. For multiple returns in close proximity and within the resolution of the outgoing laser pulse, the k^{th} return pulse will not only merge with the first return pulse shape but also with the SPAD artefacts introduced into the signal. The returned signal strength is influenced further by surface reflectivity, shape, angle of incidence and atmospheric conditions. Hence the underlying instrumental response for a single return, $\rho \in \mathbb{R}_+^p$, will have significant variability in amplitude, Γ and instrumental variation of the system. Thus a practical instrument response function (IRF) can be described as

$$\mathbf{r} = (\Gamma \cdot \delta) \star \rho, \tag{6}$$

where $\star$ is the discrete convolution operator.

B. Waveform Simulation

To train CNNs a significant amount of training data is required, ideally with full labelling for classification problems. Fully labelled data will be called ground truth (GT) for the remainder of the paper. We have developed simulation tools which enable us to generate waveforms in volume with enough variance to effectively train a neural network and apply the resulting networks to unseen real data. For a controlled experiment such as a two peak super-resolution experiment as presented in [10], the parameters are well defined and a real IRF can be obtained. In absence of a real system, it can be readily modelled [2]. It contains the aforementioned system characteristics and defines the minimum system resolution. We use a system IRF and vary the amplitude randomly for validation and a modelled IRF varied by synthetic scene information for the simulated experiment. The remainder of the data generation is straightforward and replicates a normal experimental methodology, thus for each measurement:

- Place up to K targets with amplitude at their respective position, $\mathbf{g} = [(g_0, d_0) \dots (g_{K-1}, d_{K-1})]$
- Generate for each target distance, d_k, a binary histogram, $\mathbf{h}_k = [0 \dots 1 \dots 0]$
- Convolve each $\mathbf{h}_k$ with the scaled IRF to get the individual photon count rate distribution, $\mathbf{c}_k = (g_k \mathbf{r}) \star \mathbf{h}_k$
- Add all count rate distributions to generate the full count rate distribution, $\mathbf{c}_K = \sum_{k=0}^{K-1} (\mathbf{c}_k)$

- Let β be the mean photon count rate for background noise.

This forms the resultant waveform drawn from a Poisson distribution $\phi(\lambda) \in \mathbb{N}_0^p$, with variance λ,

$$\mathbf{f}_K = \phi(\mathbf{c}_K + \beta). \tag{7}$$

For long range applications we leverage synthetic datasets [14], [15] which have rich labelled scenes with underlying exact depth ground truth. To simulate the beam expansion along a long range we employ a spatial down sampling routine which emulates the reception of multiple returns from an area within the original depth map.

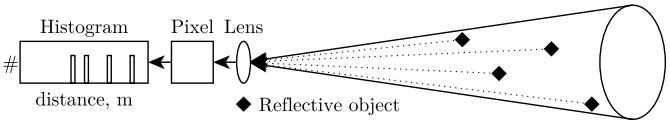


Fig. 2. Re-sampling approach for multiple returns.

While loosing the true original spatial location, the depth of all objects in the scene is fully retained. By defining the size of the sample region, $n \times n$, we define the maximum permissible number of true returns as n^2. In this paper we set $n = 3$ and therefore allow for a total of $K = 9$ returns. We further use the semantic information from the synthetic dataset to generate generic reflectivity values for labelled objects in the scene and scale them by their brightness determined by their RGB image (total power). This gives a rudimentary estimation of reflectivity and introduces variability into the training datasets. This allows us to generate histograms from varying environments with known ground truth and allows meaningful analysis of the DL networks outputs.

IV. Deep Learning & Validation

A. Network Architecture

TABLE I
NETWORK PARAMETERS FOR SUPER-RESOLUTION AND AUTOMOTIVE LONG RANGE VARIABLE NOISE LIDARNET CONFIGURATIONS

	SuperRes		Automotive		Activation
	L, W	#params	L, W	#params	
EB1	64, 64	4160	96, 48	4704	ReLU
EB2	64, 32	131136	96, 48	442464	ReLU
EB3	-	-	64, 24	147520	ReLU
Conv1D	32, 32	65568	32, 24	49184	ReLU
Conv1D	16, 32	16400	16, 24	12304	ReLU
	C	#params	C	#params	
Dense	128	2097280	256	3842304	ReLU
Dense	4096	528384	7500	1927500	SoftMax
total		2842928		6425980	

Deep learning (DL) has been effectively applied for removing noise from multi-dimensional data including 2D and 1D data with extensive work in classification problems with 100s of final possible outcomes. Our hypothesis is based on the ideal signal sparsity of LiDAR waveforms in (1) and methods allowing reduction of dimensionality of data from auto-encoder architectures.

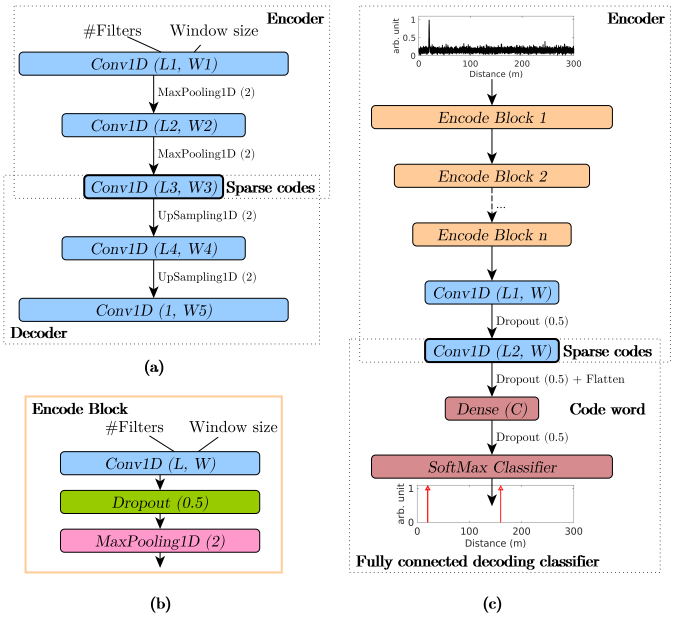


Fig. 3. LiDARNet: Taking inspiration from a convolutional auto-encoder (a), we derive an encoding block (b). The encode block reduces dimensionality and takes the parameters, L, for hidden layers i.e. expected features and the window size, W, for the kernel function. We construct an encoding stage in (c) to reduce the data to sparse codes and then classify the data using an expanding fully connected network with a code parser with C weights.

Our goal is to convert a dense LiDAR waveform to its principal peak location components as a classification problem with confidence values attached to the potential locations of objects. Starting with a convolutional auto-encoder, Fig. 3(a), we had good success on simulated data only, but could not transfer the simulation results to real data. This led to a hybrid approach, LiDARNet in Fig. 3(c), where the encoding stage of n blocks (Fig. 3(b)) is followed by two convolutional layers and combined with a traditional fully connected classification block. We provide parameters for both experiments in Table I. Both network configurations were compiled with an Adamax optimizer and a binary cross-entropy loss function suitable for classification problems. One interesting aspect of this work is that we allow for thousands of final output locations depending on the desired depth resolution either matching the input resolution or potentially enabling super-resolution.

To validate our approach, we replicate an experiment first presented in [10], where one reflector is moved away from another and is illuminated with a laser source, with a pulse width corresponding to $4\,\mathrm{cm}$, which can not resolve very close separations despite a sampling resolution of $1\,\mathrm{mm}$. For this case we set $K = 2$. We generate 25000 synthetic waveforms with random separations, amplitudes and noise within the experimental parameters and train our LiDARNet in the SuperRes configuration with binary ground truth labels.

B. Validation

We expect traceable features which match the signal model and system properties. As shown in Figure 4, the CNN kernels extract recognisable features which are discarded as the dimensionality is reduced or propagated if significant. The

first CNN layer recognises the SPAD tail as a feature, the region of interest where peaks are present or sections the signal into before and after the SPAD tail. The region of interest and sectioned signal parts are propagated deeper into the network while SPAD tails and noise features are rejected. The signal area is then sectioned until generally 2 spikes are observed as outputs in a low dimensional feature vector for each respective CNN kernel. From those sparse codes a code word is generated, which is then used to decode the peak locations. This aligns well with the signal model in (1).

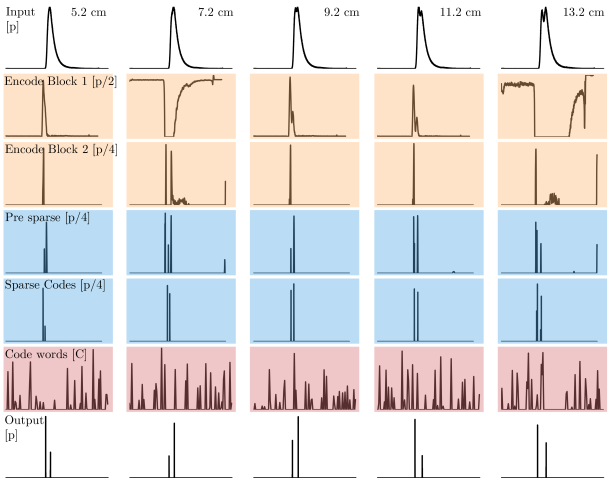


Fig. 4. Top to bottom: Selection of two-peak super-resolution inputs; illustrative selection of first encoding layer activations for each input, showing a variety of traceable features, such as region of interest, system characteristics, and step edge; second encoding layer forming proposals; enforcing proposals from CNN filters; and sparse-code proposals; code words for each input; and final outputs.

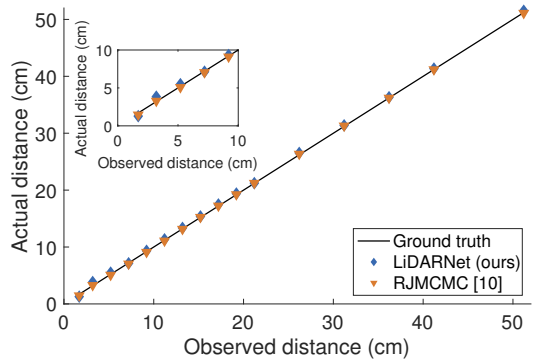


Fig. 5. Using our purely synthetic data for training and applying the resultant network to unseen real data from the super-resolution benchmark, we get comparable results to the traditional signal processing method.

We apply this network to real data from [10] and evaluate it in Table II and compare the results visually in Figure 5. Our data driven approach performs comparably to the statistical approach. Therefore, we can demonstrate the ability to perform super-resolution on real-data with only simulated synthetic training data. This validates our training approach.

V. Results

Both network configurations were implemented using Keras and Tensorflow in Python 3.6 on a dual Intel Xeon E5-2630v3 (2.4 GHz) with 48 GB of memory and a NVidia RTX 2080 Ti (12 GB).

TABLE II
Evaluation of a real super-resolution benchmark using our synthetically trained LiDARNet

	RJMCMC [10]		LiDARNet (ours)	
ground truth (cm)	Mean (cm)	error (cm)	Mean (cm)	error (cm)
1.7	1.462	0.238	1.281	0.419
3.2	3.281	0.081	3.843	0.643
5.2	5.086	0.114	5.489	0.289
7.2	7.053	0.147	7.136	0.064
9.2	9.108	0.092	9.332	0.132
11.2	11.092	0.108	11.345	0.145
13.2	13.155	0.045	13.357	0.157

The following experiment is designed to demonstrate that the proposed network architecture in the automotive configuration is capable of performing super-resolution as well as effectively identify surfaces at long range in noise simultaneously. We use 100000 waveforms generated from synthetic scenes [14], [15] for training and an additional 4000 waveforms for testing, with $K = \{0, 9\}$, a mean return rate of 19.4 ($g = \{0, 278.4\}$) and background noise levels randomly varying for $\beta = \{0.04, 38.3\}$. The network is trained for 150 epochs for each waveform, normalized by maximum bin value, together with the corresponding binary ground truth labels.

First, we evaluate the receiver operating characteristic (ROC) using the test dataset with 4000 waveforms. We allow for ± 12 cm with a successful bin classification if greater than the threshold $t = \{0.001, 0.01\}$. We see a good true-positive rate at above 95% as the threshold increases. The overall accuracy is excellent at above 99% throughout.

Second, to evaluate the multi-return reconstruction capability of the network, we compare our network output convolved with r to waveform reconstructions from reverse jump Markov chain Monte Carlo (RJMCMC) [10] with min-max normalized amplitude. Neither method requires prior knowledge of the exact number of peaks present in each wave. We evaluate on 1000 test waveforms with peak signal-to-noise ratio (PSNR) as well as mean-squared error (MSE) values shown in Table III and show an illustrative selection of results in Figure 7 for $t = 0.04$ versus their normalized ideal waveform $\mathbf{c}_K$. We note that our network achieves a consistent increase in PSNR with an average of above 43 dB slightly better than RJMCMC for a mean input PSNR of 8.25 dB. Further, RJMCMC requires more than 5 s of runtime per waveform, while LiDARNet processes a single waveform in 3.4 ms with excellent parallelism e.g. 64 waveforms in 43 ms.

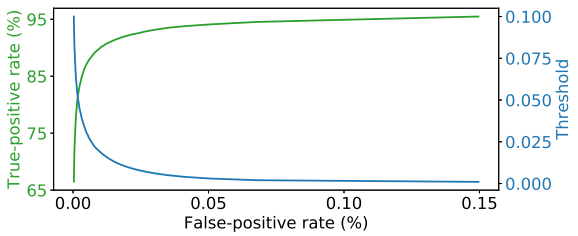
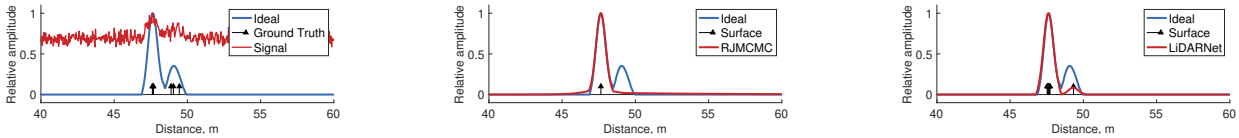


Fig. 6. Receiver operating characteristics (ROC) for automotive LiDARNet with a mean accuracy of 99.98%.



(a) Reference 3.15 dB (b) RJMCMC [10] 36.05 dB (c) **LiDARNet (ours)** 37.12 dB

(d) Reference 3.79 dB (e) RJMCMC [10] 31.59 dB (f) **LiDARNet (ours)** 40.37 dB

Fig. 7. Waveform reconstructions and detected surface returns for noisy long-range LiDAR waveforms – Top row: (a) two clusters of multiple returns, (b) RJMCMC recovers a single cluster, (c) while LiDARNet retrieves both – Bottom row: (d) several closely resolved surfaces, (e) RJMCMC only recovers one cluster correctly, while (f) LiDARNet recovers them with multiple returns. Ground truth and surface positions scaled for visual clarity.

TABLE III
EVALUATION OF LIDARNET IN AUTOMOTIVE CONFIGURATION FOR 1000 SIMULATED LIDAR WAVEFORMS COMPARED TO IDEAL WAVEFORMS.

	Signal	RJMCMC [10]	LiDARNet
PSNR [dB]	8.25	40.43	43.50
MSE	0.2935	0.0008	0.0006
time [ms]	-	> 5000	3.4

VI. CONCLUSION

We have presented the application of deep learning methods to process full-waveform LiDAR signals with multiple returns with a training approach which relies primarily on simulated data but has been validated on a real data benchmark.

Our validation experiment demonstrates the capability of a machine learning classifier to perform peak location extraction from high resolution LiDAR waveforms where two target signatures are separated by various distances ranging from super-resolution peak extraction, to resolving two close but separated return surfaces. We train our classifier on *simulated* data only and validate on unseen *real* data with good results.

Another experiment demonstrates this data driven approach applied to long range waveform LiDAR with a high dynamic range. The simulated dataset has varying number of returns and a wide range of background noise levels. LiDARNet is capable of identifying significant peak locations and significant clusters directly with better signal reconstruction performance than a statistical approach while also being several orders of magnitudes faster and massively parallel.

This enables a straightforward fast point cloud generation for high resolution arrayed LiDAR systems to allow further processing on the extracted 3D data, while retaining multiple return information.

REFERENCES

[1] Ayush Bhandari, Andrew M Wallace, and Ramesh Raskar, "Super-resolved time-of-flight sensing via FRI sampling theory," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 4009–4013, IEEE.

[2] Sara Pellegrini, Gerald S Buller, Jason M Smith, Andrew M Wallace, and Sergio Cova, "Laser-based distance measurement using picosecond resolution time-correlated single-photon counting," *Meas. Sci. Technol*, vol. 11, no. 00, pp. 712–716, 2000.

[3] Sarrah. M. Patanwala, Istvan Gyongy, Neale A. W. Dutton, Bruce. R. Rae, and Robert. K. Henderson, "A Reconfigurable 40nm CMOS SPAD Array for LiDAR Receiver Validation," in *International Image Sensor Workshop*, 2019.

[4] Preethi Padmanabhan, Chao Zhang, and Edoardo Charbon, "Modeling and Analysis of a Direct Time-of-Flight Sensor Architecture for LiDAR Applications," *Sensors*, vol. 19, no. 24, pp. 5464, dec 2019.

[5] Geoffrey E. Hinton and Ruslan R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, jul 2006.

[6] Corneliu T.C. Arsene, Richard Hankins, and Hujun Yin, "Deep Learning Models for Denoising ECG Signals," in *2019 27th European Signal Processing Conference (EUSIPCO)*. sep 2019, pp. 1–5, IEEE.

[7] Antonia Creswell and Anil Anthony Bharath, "Denoising Adversarial Autoencoders," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 968–984, apr 2019.

[8] David L. Donoho and Iain M. Johnstone, "Threshold selection for wavelet shrinkage of noisy data," *Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, pp. A24—-A25, 1994.

[9] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, aug 2006.

[10] Sergio Hernández-Marín, Andrew M. Wallace, and Gavin J. Gibson, "Bayesian analysis of lidar signals with multiple returns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2170–2180, 2007.

[11] Yoann Altmann, Stephen McLaughlin, Miles J. Padgett, Vivek K. Goyal, Alfred O. Hero, and Daniele Faccio, "Quantum-inspired computational imaging," 2018.

[12] Gangping Liu and Jun Ke, "Deep-learning for super-resolution full-waveform lidar," in *Optoelectronic Imaging and Multimedia Technology VI*, Qionghai Dai, Tsutomu Shimura, and Zhenrong Zheng, Eds. nov 2019, vol. 11187, p. 38, SPIE.

[13] Andreas Aßmann, Brian Stewart, João F.C. Mota, and Andrew M. Wallace, "Compressive Super-Pixel LiDAR for High-Framerate 3D Depth Imaging," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Ottawa, Canada, 2019.

[14] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig, "Virtual Worlds as Proxy for Multi-Object Tracking Analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4340–4349.

[15] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3234–3243.